"All You Can Eat" Ontology-building: Feeding Wikipedia to Cyc

Catherine Legg
University of Waikato
22 April, 2009

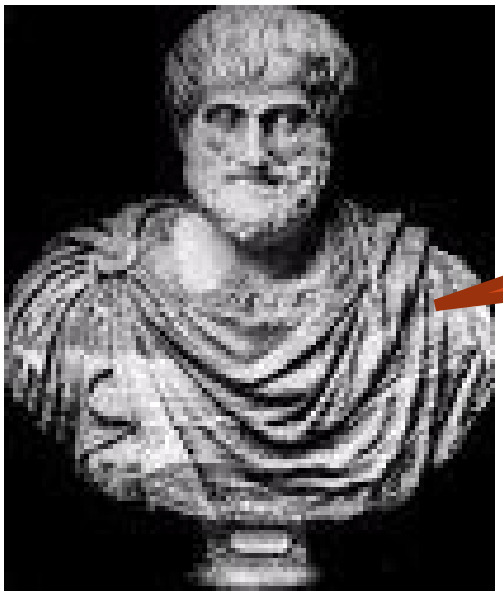
Agenda

1. What is Ontology?
2. The Cyc Ontology
3. Wikipedia
4. Automated Ontology Integration

Agenda

1. **What is Ontology?**
2. The Cyc Ontology
3. Wikipedia
4. Automated Ontology Integration

The philosophical discipline of **ontology** was invented by Aristotle:



όντος: (being) +

λόγος: (theory of)

It was called by him ‘first philosophy’ (i.e. the fundamental science)

It builds a theory of all the different **kinds** of things ('categories') that exist in reality

Examples might include:

Physical objects

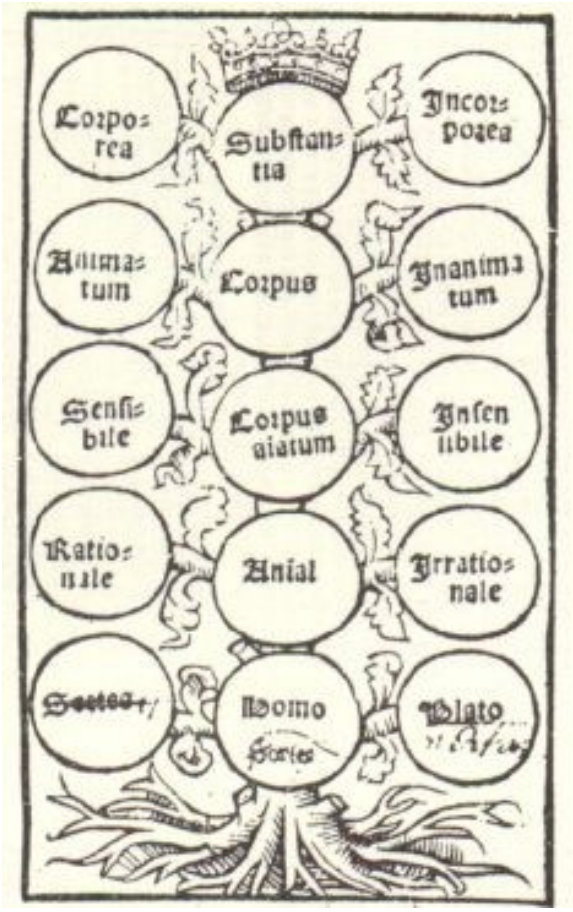
Times

Events

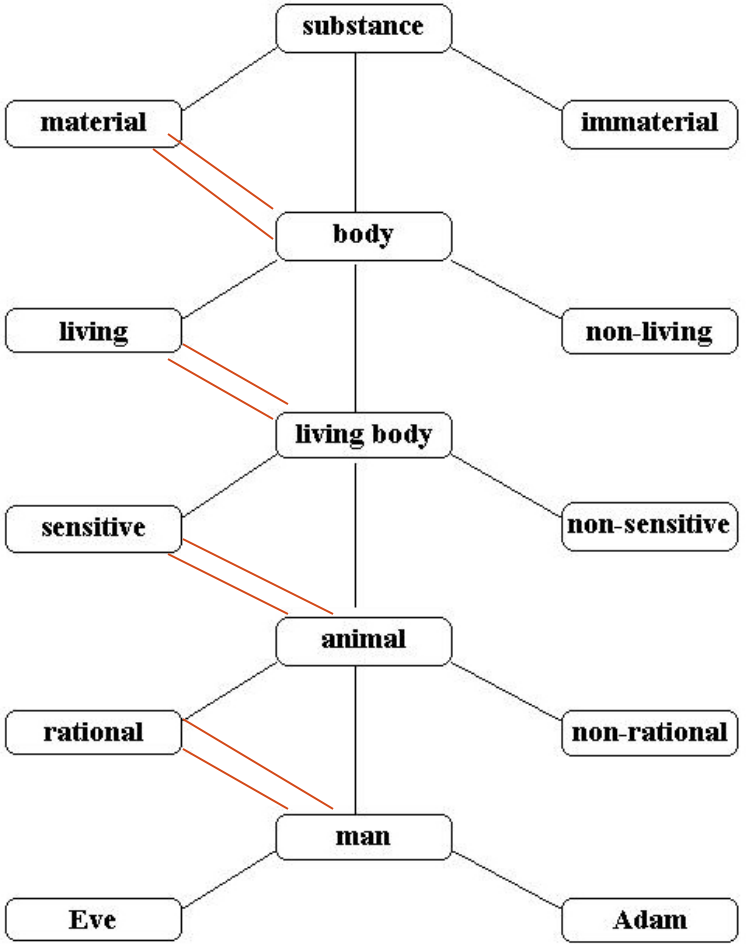
Relationships

Numbers

Traditionally ontologies were built into a taxonomic structure, sometimes referred to as a ‘tree of knowledge’:



‘Tree of Porphyry’ (3rd century A.D.)



Why would a **software engineer** build an ontology?

Believe it or not, many problems arise when dealing with information, to do with understanding what **kinds** of things are being referred to.

Can anyone think of any possible examples?

What kind of a thing is an **address**? We can treat it as just a string of symbols: "**`7BrettonTerraceHillcrest3216`**".

But we can parse, manipulate and use addresses much more intelligently if we understand that they are made up of a reference to a **building number**, a **street** and a **suburb.**

Ontologies also seem to be needed for consistency-checking information.

For instance, what is wrong with the following lines of a human resources database?

XYZCo ID #	birth date	hire date	salu-tation	first name	last name	emerg contact	signif other
8041	9/1/57	8/5/91	Mr	Pat	Jones	8053	8053
8053	3/3/49	2/9/48	Ms	Jan	Smith	8053	8199

Agenda

1. What is Ontology?
2. **The Cyc Ontology**
3. Wikipedia
4. Automated Ontology Integration

Brief History of the Cyc Project

- **c. 1967** – Artificial intelligence (AI) is used on toy problems.
- **c. 1977** – 'Expert systems' reason in narrow domains, but break down when asked to consider new problems ('brittleness')
- **c. 1983** – Key AI researchers (e.g. Marvin Minsky) decide that to make any further progress, computers have to have "common-sense".
- **1984** – The Cyc project is begun at MCC, to capture that common-sense in a giant ontology.
- **1994** – The company Cycorp is formed, to continue the project.
- **2002** – OpenCyc released (www.opencyc.org) …

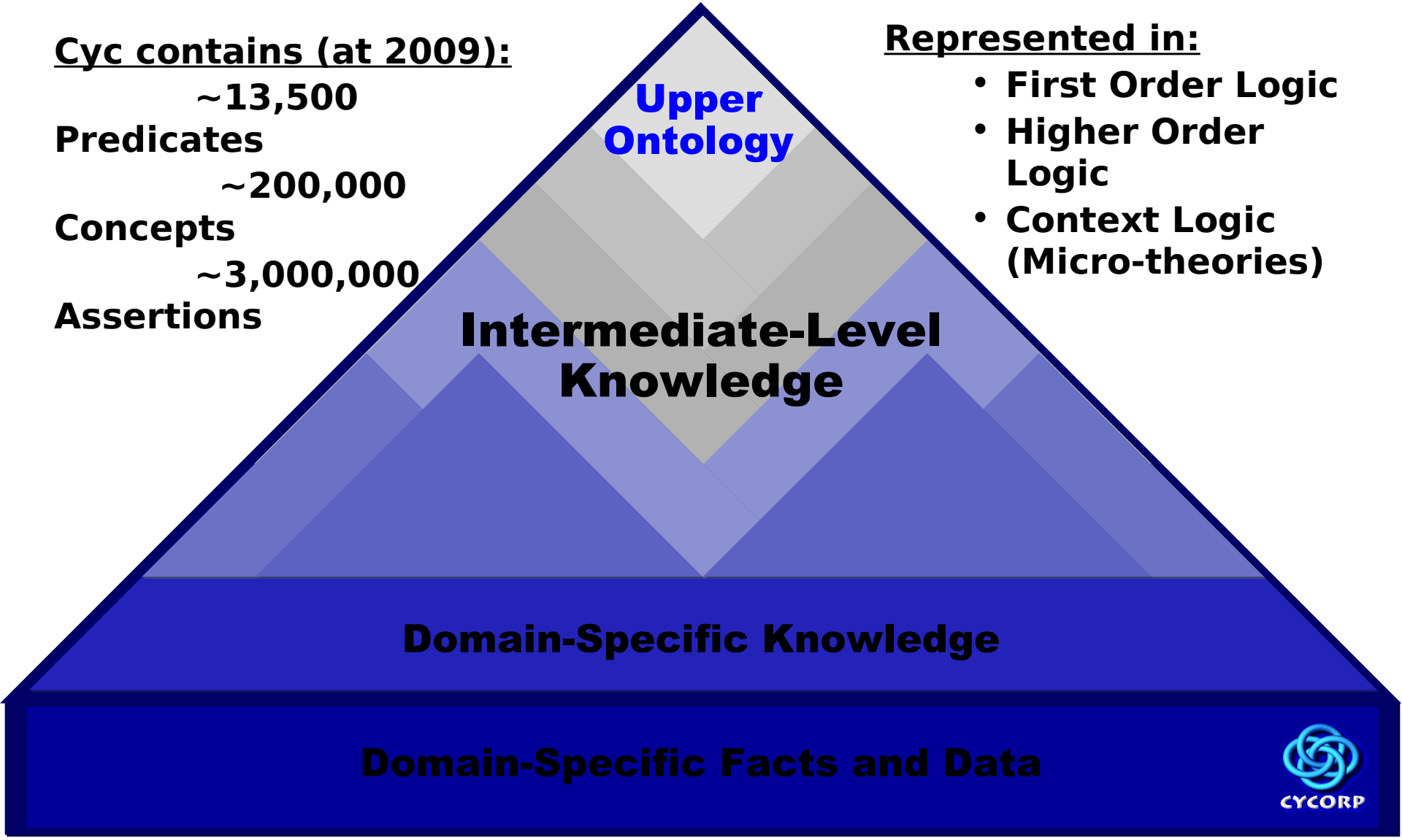
Cyc Ontology & Knowledge Base

Cyc contains (at 2009):

~13,500 Predicates

~200,000 Concepts

~3,000,000 Assertions



Represented in:

- **First Order Logic**
- **Higher Order Logic**
- **Context Logic (Micro-theories)**

CYCORP

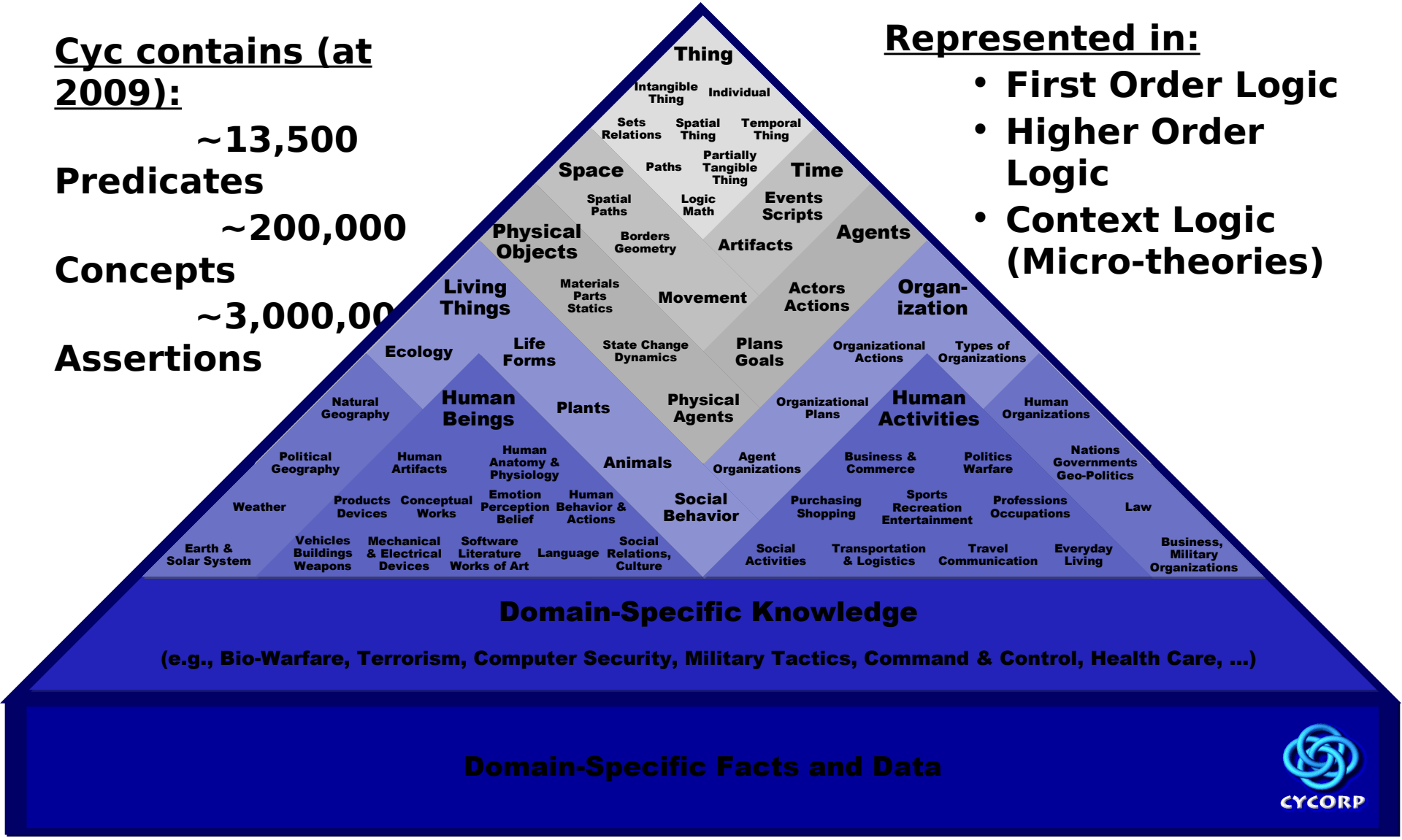
Cyc Ontology & Knowledge Base

Cyc contains (at 2009):

~13,500 Predicates

~200,000 Concepts

~3,000,00 Assertions

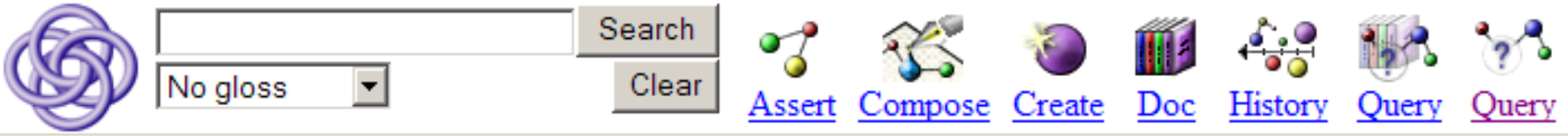


Represented in:

- **First Order Logic**
- **Higher Order Logic**
- **Context Logic (Micro-theories)**

CYCORP

Cyc ‘inference engine’



Query Tool

Go to section : [New Inference] [Inference Parameters]

Hide **Focal Inference** **Actions :** [Debug] [Examine] [Defocus] [Destroy] [Template OE] [Save As Query] [Sa
Test] [SubL Query] [Query Graph]

Mt : (MtSpace CurrentWorldDataCollectorMt-NonHomocentric (MtTimeDimFn Now))
EL Query :
(not
(isa BillGates ParkingMeter))

Common-sense knowledge

Status : Suspended, Exhaust Total

Query was proven True [Explain]

Proven true

Typical workday of a Cycorp 'ontological engineer'

Key stages in ontological design

1. Get clear on what questions you want the system to answer, and how.

2. Identify already existing concepts and assertions in the ontology that can be re-used

3. Create new concepts and make assertions on them

4. [*Crucial*] Test the new assertions using the inference engine. Does the system give you the answers you need? If not, debug and restart 4.

E.g. I might want the system to know…

- There is a University called “Melbourne University”.
- Melbourne University is in Melbourne.
- It has many academic departments.
- It owns buildings.
- It teaches students.
- It is an institution.

…and so on

Collection : University

This concept already exists

GAF Arg : 1

Mt : UniversalVocabularyMt
isa : Collection
not isa : AtemporalNecessarilyEssentialCollectionType

Mt : BaseKB
isa : PublicConstant

Mt : BookkeepingMt
isa : PublicConstant-DefinitionalGAFsOK PublicConstant-CommentOK

Mt : AcademicOrganizationVocabularyMt
isa : EELDNaturalDataConstant ExistingObjectType
genls : GeographicalAgent DegreeGrantingHigherEducationInstitution ResearchOrganization

Mt : AcademicOrganizationMt
disjointWith : College SchoolInUniversity-DegreeGranting CollegeInUniversity-DegreeGranting

Mt : AcademicOrganizationVocabularyMt
comment : "A specialization of DegreeGrantingHigherEducationInstitution, EducationalOrganization, ResearchOrganization, and GeographicalAgent. Instances of University are educational organizations (whose official names usually, but not always, include the word 'University') at which university-level teaching and research takes place. Some universities comprise multiple "colleges" and professional schools. Note that this collection includes universities that lie within a more or less local and spatially contiguous campus as well as those that consist of a system of such campuses. Thus both UniversityOfTexasAtAustin and its encompassing system (see subOrganizations) UniversityOfTexas are instances of University."

Mt : CycSubjectClumpsMt
cycSubjectClumps : OrganizationsByFunction-Organization-CSC

Mt : BaseKB
definingMt : AcademicOrganizationVocabularyMt

Create Constant

Cancel Create | Clear | Create Constant

Enter name for new constant :

Complete | MelbourneUniversity

--For detailed help and warnings about creating, please read the Cyc Naming Conventions documentation.

I now make a new concept, which I will categorise ‘under’ *University*

Constant : MelbourneUniversity

Bookkeeping Assertions :

(myCreator MelbourneUniversity CathyLegg) in BookkeepingMt
(myCreationPurpose MelbourneUniversity CycSecure) in BookkeepingMt
(myCreationTime MelbourneUniversity 20021206) in BookkeepingMt
(myCreationSecond MelbourneUniversity 225222) in BookkeepingMt

GAF Arg : 1

Mt : UniversalVocabularyMt
isa : Entity

Mt : UniversityDataMt
isa : University

This is the key assertion

Mt : OrganizationDataMt
comment : "The University of Melbourne, located in leafy Parkville, Melbourne, Australia."
foundingDate : (YearFn 1854)

Mt : CyclistsMt
hasAlumni : CathyLegg

Mt : EnglishMt
nameString : "Melbourne Uni"
preferredNameString : "the University of Melbourne"

Mt : UniversityDataMt
residenceOfOrganization : CityOfMelbourneAustralia

Mt : OrganizationDataMt
subOrganizations : UniversityHouse

Literal Query

Now see how much ontological knowledge is 'inherited' onto my new concept *Melbourne University*, just by placing it under the category of *University*

Last query in EverythingPSC :
(isa MelbourneUniversity ?ARG2)

55 answers for ?ARG2 :

AcademicOrganization Agent Agent-Generic Agent-Underspecified
Artifact-Generic Boundary-Underspecified
CompositeTangibleAndIntangibleObject Container-Underspecified
CulturalThing DegreeGrantingHigherEducationInstitution EducationalOrganization
Entity Expression-Underspecified FunctionalSystem GeographicalAgent
GeographicalRegion GeographicalThing Group HigherEducationInstitution
InanimateThing InanimateThing-NonNatural Individual InformationStore
IntelligentAgent Landmark-Underspecified Location-Underspecified MultiIndividualAgent
Organization PartiallyIntangible PartiallyIntangibleIndividual PartiallyTangible
Place Region-Underspecified ResearchOrganization SocialBeing SomethingExisting
SpaceRegionLimit SpatialThing SpatialThing-Localized Surface-Generic
Surface-Open Surface-Physical SurfaceRegion-Underspecified System-Generic TemporalThing
Thing Trajector-Underspecified TwoOrHigherDimensionalThing
University
(CollectionUnionFn (TheSet Agent-Generic (GroupFn Agent-Generic))) (CollectionUnionFn (TheSet Informa

Agenda

1. What is Ontology?
2. The Cyc Ontology
3. **Wikipedia**
4. Automated Ontology Integration

Wikipedia is astounding! (at 2009):

10M articles in **250 different languages**

2.4M articles in the English version, referred to by **3M different terms**

~25 hyperlinks per article

400 000 categories in the English version, with an average of **19 articles and 2 subcategories** in each

175 000 templates for semi-structured data-entry (including **9 000 'infoboxes')**

full editing history for every article is recorded

a **discussion page** for every article

and all for……..

Wikipedia as an ontology

- **articles** can be viewed as basic concepts
- **infoboxes** can be mined for facts about those concepts
- **hyperlinks** between articles can be mined to determine ‘semantic relatedness’ between concepts
- **categories** organise the articles into conceptual groupings. Although it must be said that Wikipedia categories are far from a Cyc-style principled taxonomy enabling knowledge inheritance

For example, consider the following category….

Pages in category "Pork"

The following 37 pages are in this category, out of 37 total. This list may not re ecent changes (learn more).

- Pork

*

- Religious restrictions on the consumption of pork

B

- Back bacon
- Bakkwa

C

- Capicola
- Charcuterie
- Chicken fried bacon
- Chitterlings
- Chocolate covered bacon
- Ciccioli
- Crépinette

D

- Domestic pig

F

- Full breakfast

K

- Kassler

L

- Lardo
- Lean Hog
- Lechón
- Lomo (food)

M

- Mett

P

- PSE meat
- Pancetta
- Pickled pigs feet
- Pig Candy
- Pig pickin'
- Pig roast

P cont.

- Pork jelly
- Pork rind
- Pork roll

R

- Ractopamine
- Rostbrätel
- Rousong

S

- Salt pork
- Scottish pork taboo
- Suckling pig

T

- Tasso ham
- Tocino

V

- Valle d'Aosta Lard d'Arnad

Agenda

1. What is Ontology?
2. The Cyc Ontology
3. Wikipedia
4. **Automated Ontology Integration**

Current work in the Digital Libraries Lab, University of Waikato, is focusing on mining the vast quantity of data which exists in Wikipedia and adding it to the more structured taxonomy that exists in Cyc. We like to call this project……..

Feeding Wikipedia to Cyc

The work built on earlier work by Olena Medelyan and Cathy Legg (presented at Chicago AAAI, 2008), finding mappings between equivalent concepts in Cyc and Wikipedia.

Method	Cyc terms	Percent mapped
Total terms available	163,000	
Common sense terms	83,900	
Exact (1-1) mappings	33,500	40%
Further mappings after disambiguation(2 ways)	8,800	10%
Total mapped	**42,300**	**50%**

The mapping algorithm used there was adapted and improved as follows...

Stage A&B: easy cases where 1-1 matches are identifiable using either title strings or synonyms

Exact mappings

via Cyc synonyms

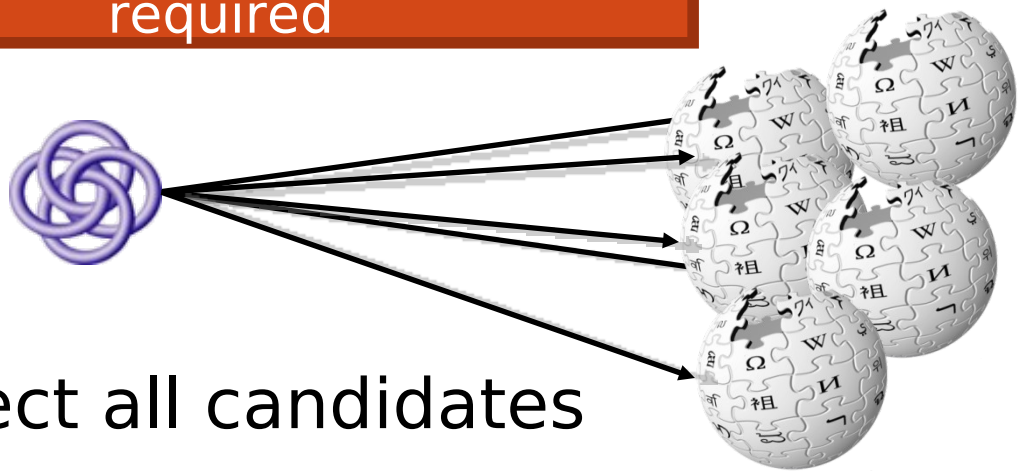
CityOfSaarbrucken → Saarbrücken

Get synonyms asserted in Cyc, e.g. using #$nameString

Use Wikipedia redirects

Saarbrucken —matches→ Saarbrucken

Stage C: a number of possible candidates on the Wikipedia side, semantic disambiguation required



- Collect all candidates
- Compute commonness of each candidate
- Collect context from Cyc
- Compute similarity to context

Collecting candidates

Kiwi

matches

- Kiwi (people)
- Kiwi (bird) — redirects → Kiwi
- Kiwi (fruit) — redirects → Kiwifruit
- Kiwi (disambiguation) — disambiguates → Kiwis (rugby league)
  - New Zealand national rugby league team

Collect context from Cyc

Kiwi 

isa FlightlessBird → Flightless bird

isa BiologicalSpecies → Species

isa Animal → Animal

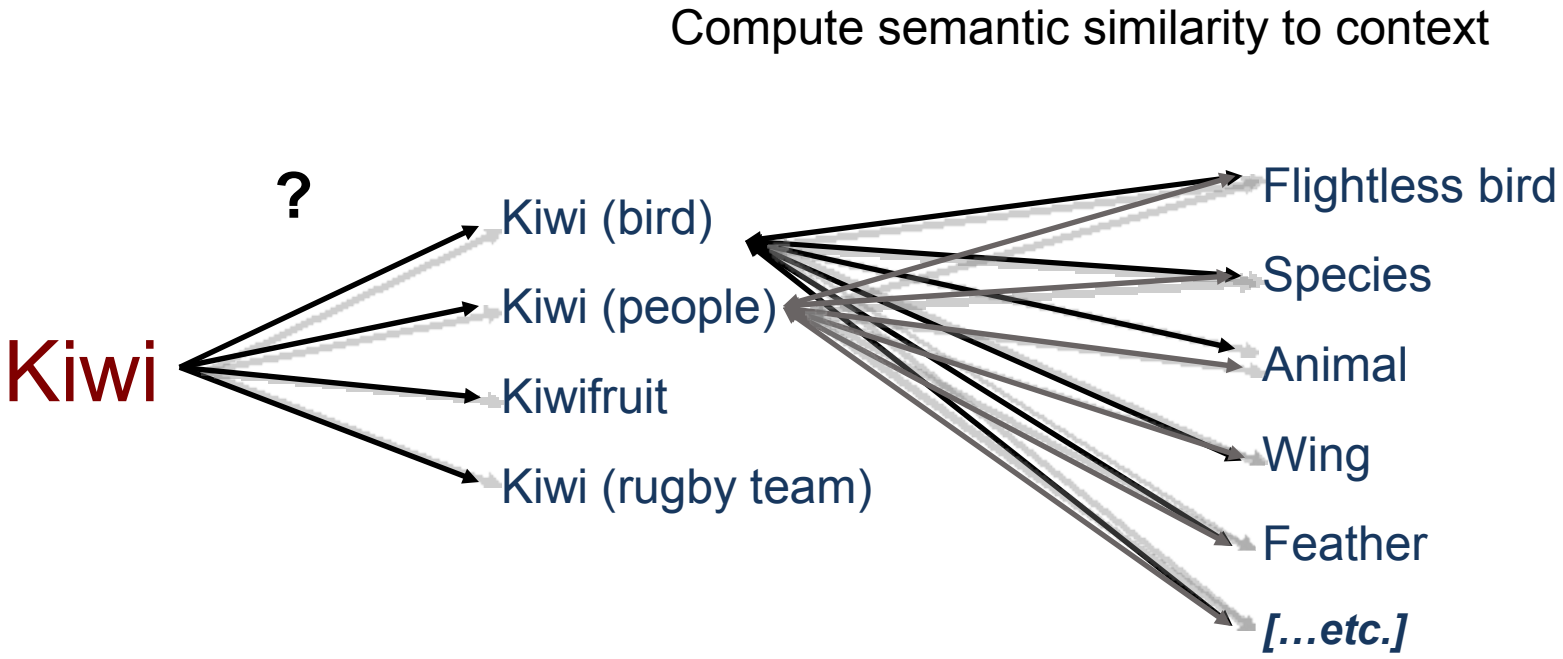
conceptuallyRelated Wing-AnimalBodyPart → Wing

conceptuallyRelated Feather → Feather

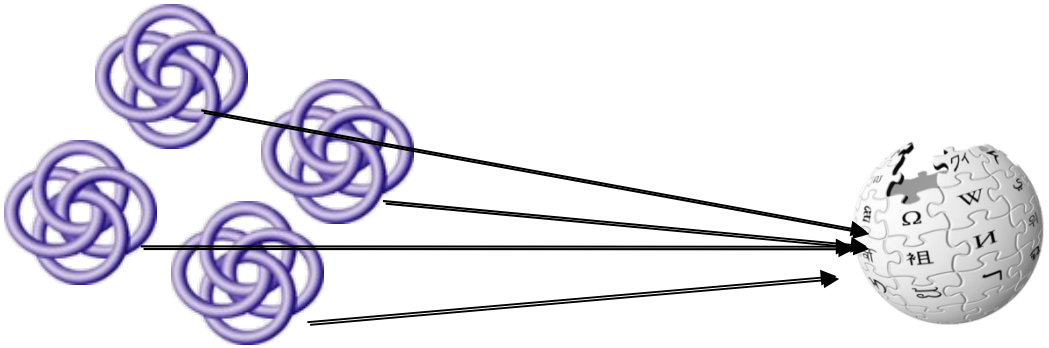
conceptuallyRelated BirdFood → ?

The Cyc terms come from the immediately surrounding ontology, and the Wikipedia mappings come from Stages A and B. Cyc context terms that have no mapping to Wikipedia yet are ignored.

Determine best candidate



Stage D: reverse disambiguation (many Cyc concepts map to the same Wikipedia article)



In this stage we eliminate many candidate mappings by attempting to map back from the Wikipedia article to the Cyc term, and discarding mappings which don't pass this test. For example the term #$DirectorOfOrganisation incorrectly maps to the article Film director, but when we attempt to find a Cyc term from Film director we get #$Director-Film.

This reduces the number of mappings by 43%, but increases precision considerably.

Current work (Sarjant, Robinson and Legg) builds on the mappings by finding new concepts in Wikipedia and adding them to Cyc

First we find mapped concepts where the Wikipedia article has an equivalent category (about 20% of mapped concepts). E.g. the article *Israeli Settlement* has an equivalent category *Israeli Settlements*:

Pages in category "Israeli settlements"

The following 72 pages are in this category, out of 72 total. This list may not reflect recent changes (learn more).

- Israeli settlement
- Yesha

A

- Adei Ad
- Almon, Mateh Binyamin
- Argaman
- Asa'el

B

- Barkan Industrial Park
- Beit Aryeh-Ofarim
- Beka'ot

D

- Dolev

H cont.

- Havat Gilad
- Havat Skali
- Havat Yair
- Hermesh
- Hilltop 26
- Hinanit

K

- Kalya
- Kfar Adumim
- Kfar Eldad
- Kfar HaOranim

L

- Livne

N cont.

- Netiv HaGdud
- Neve Daniel
- Neveh Erez
- Niran
- Nofei Prat
- Nokdim

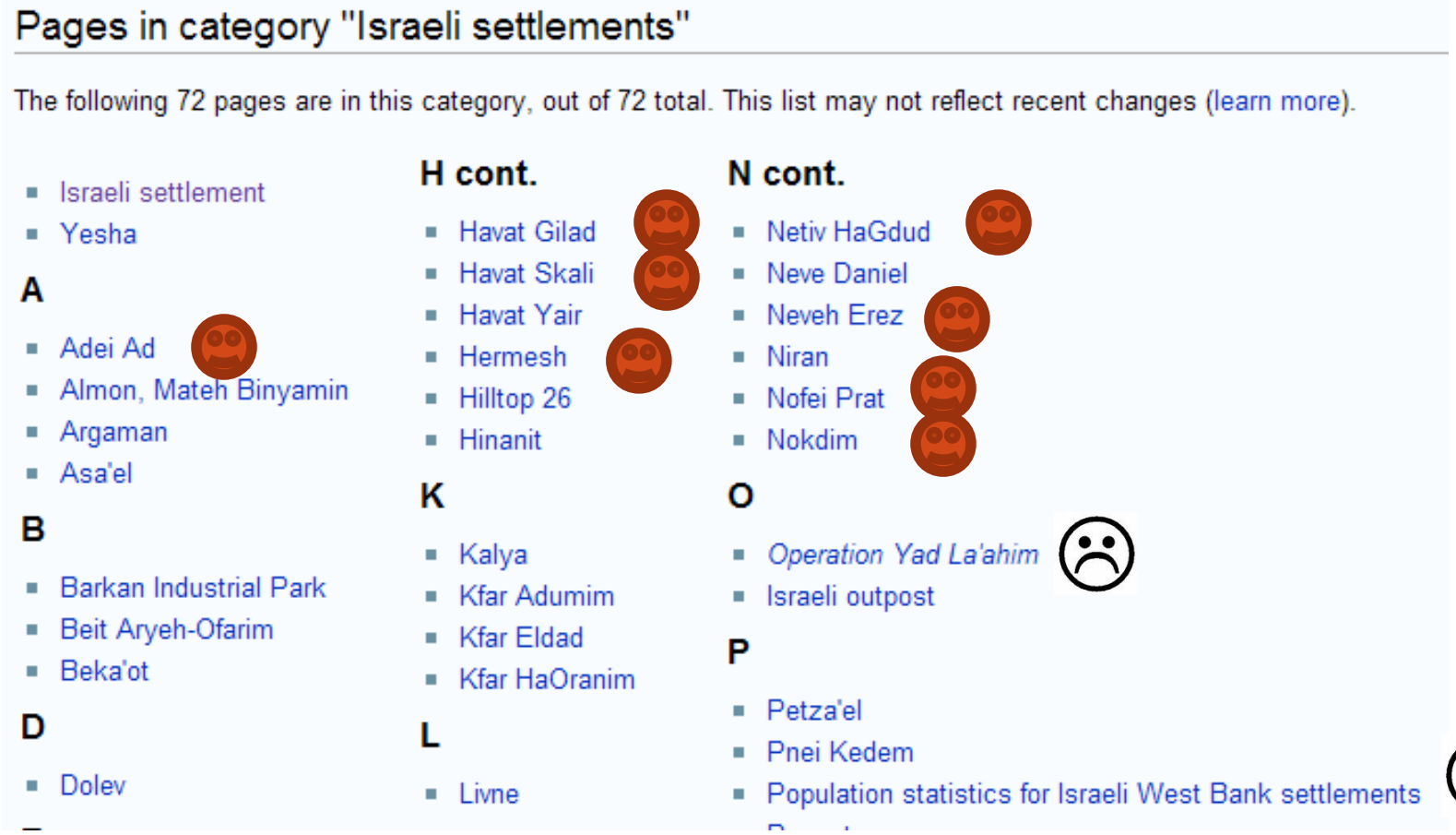
O

- *Operation Yad La'ahim*
- Israeli outpost

P

- Petza'el
- Pnei Kedem
- Population statistics for Israeli West Bank settlements

We then mine this category for new concepts which belong under the mapped Cyc concept, according to the Cyc taxonomy. For instance:



Pages in category "Israeli settlements"

The following 72 pages are in this category, out of 72 total. This list may not reflect recent changes (learn more).

- Israeli settlement
- Yesha

A

- Adei Ad
- Almon, Mateh Binyamin
- Argaman
- Asa'el

B

- Barkan Industrial Park
- Beit Aryeh-Ofarim
- Beka'ot

D

- Dolev

H cont.

- Havat Gilad
- Havat Skali
- Havat Yair
- Hermesh
- Hilltop 26
- Hinanit

K

- Kalya
- Kfar Adumim
- Kfar Eldad
- Kfar HaOranim

L

- Livne

N cont.

- Netiv HaGdud
- Neve Daniel
- Neveh Erez
- Niran
- Nofei Prat
- Nokdim

O

- *Operation Yad La'ahim*
- Israeli outpost

P

- Petza'el
- Pnei Kedem
- Population statistics for Israeli West Bank settlements

We call these '<u>true children</u>'.

We identify true children by:

We obtain approx. 20K new concepts this way.

1.Parsing the first sentences of Wikipedia articles:

Havat Gilad (Hebrew: חַוַת גִּלְעָד, *lit.* Gilad Farm) is an Israeli settlement outpost in the West Bank.

Netiv HaGdud (Hebrew: נְתִיב הַגְּדוּד, *lit.* Path of the Battalion) is a moshav and Israeli settlement in the West Bank.

Kfar Eldad (Hebrew: כפר אלדד) is an Israeli settlement and a Communal settlement in the Gush Etzion Regional Council, south of Jerusalem.

The Yad La'achim operation (Hebrew: **מבצע יד לאחים**, "Giving hand to brothers") was an operation that the IDF performed during the disengagement plan.

Table 2. The regular expressions used to parse an article's first sentence.

Regexp format	Example
X are a Y	Bloc Party are a British...
X is one of the Y	Dubai is one of the seven...
X is a Z of Y	The Ariegeois is a breed of dog...
X are the Y	The Rhinemaidens are the three...
Xs are a Y	Hornbills are a family of bird...
Xs are Y	Bees are flying insects...
The X is one of the Y	The Achaeans is one of the collective...
X is a/the Y	Batman is a fictional character...
X was/were a Y	Kipchaks were an ancient Turkic...

Where X is the candidate new child and Y is a hyperlink to a relevant Wikipedia article that has been mapped to a Cyc collection. We ended up loosening these reg exps to allow one or 2 arbitrary words around X and Y, gaining more children at little loss in accuracy.

2. "Infobox pairing" – If another article in the category shares an infobox template with 90% of true children, we include it even if its first sentence doesn't parse as a true child.

Echidna flea

From Wikipedia, the free encyclopedia

The **echidna flea** (*Bradiopsylla echidnae*) is thought to be the world's largest flea and it parasitises the short-beaked echidna. It reaches 4 millimetres in length.

References [edit]

- Australian Faunal Directory

Echidna flea	
Scientific classification	
Kingdom:	Animalia
Phylum:	Arthropoda
Class:	Insecta

We obtain approx. 15K new concepts this way.

Chigoe flea

From Wikipedia, the free encyclopedia

The **chigoe flea** or **jigger** (*Tunga penetrans*) is a parasitic arthropod found in tropical climates, especially South America and the West Indies. At 1 mm long, the chigoe flea is the smallest known flea. Breeding female chigoes burrow into exposed skin and lay eggs, causing intense irritation. After this point, the skin lesion looks like a 5 to 10 mm white spot with a central black dot, which are the flea's exposed hind legs, respiratory spiracles and reproductive organs.

If the flea is left within the skin, infection and/or other dangerous complications may ensue.

The free-living flea is a poor jumper and can only reach a height of around 20 cm; therefore the use of closed shoes (as opposed to sandals or slippers) is an effective way of preventing infection.[1]

Chigoe flea	
	
Scientific classification	
Kingdom:	Animalia
Phylum:	Arthropoda

The category which gives rise to the child we call its '<u>birth parent</u>'. However in the first sentence of its Wikipedia article we are often able to find further parents ('<u>God-parents</u>') which have been mapped to Cyc. E.g.:

Mark A. Altman

From Wikipedia, the free encyclopedia

Mark A. Altman is a film producer, screenwriter and actor. In 1998, he won Best New Writer at AFI Fest. His credits include:

- *The Specials* (producer, actor)
- *Free Enterprise* (writer, producer, actor)
- *House of the Dead* (writer, producer)
- *House of the Dead 2* (writer, producer, actor)
- *Room 6* (writer, producer, actor)
- *All Souls Day* (writer, producer)

Previously, he was a writer and editor of books and magazines relating to science fiction. Editor and co-creator, with Chris Gore, of *Sci-Fi Universe* magazine, Altman later, through his Mindfire Entertainment published *CFQ* (formerly *Cinefantastique*).

Godparent
#$Producer-Movie

Birth parent
#$Screenwriter

Godparent
#$Actor

This gives the new concepts a richer network of inter-relations (Cyc is a graph-structure not a tree structure)

Individual : MarkAAltman

Bookkeeping Assertions :

(myCreationTime MarkAAltman 20090320) in BookkeepingMt
(myCreationSecond MarkAAltman 192107) in BookkeepingMt

GAF Arg : 1

Mt : UniversalVocabularyMt
isa : Individual

Mt : WikipediaToCycDataMt
isa : Actor Screenwriter
comment : "Mark A. Altman is a film producer, screenwriter and actor."

Mt : WikipediaToCycImplementationMt
individualityStatusDeterminedBy : "Article title parsing"

A further issue we need to deal with is that Cyc enforces a principled ontological distinction between:

Individuals – e.g. **#$George-TheCat**

Collections – e.g. **#$Cat**

*So far, these distinctions have always been made by philosophically trained ontological engineers working at Cycorp. But **we** need a way of making them **automatically**…*

A further issue we need to deal with is that Cyc enforces a principled ontological distinction between:

Individuals – e.g. **#$George-TheCat**

Collections – e.g. **#$Cat**

*So far, these distinctions have always been made by philosophically trained ontological engineers working at Cycorp. But **we** need a way of making them **automatically**…*

We address this strategy via a set of overlapping heuristics:

1. Parsing the first sentence of the Wikipedia article for 'regular expressions' (e.g. '*Xs are a kind of Y*': Collection). 7% of cases

2. Patterns of capitalization in multi-word Wikipedia article titles: If later words are capitalized it is probably an Individual (e.g. *Bill Gates, American Red Cross*). If they are not it is probably a Collection (e.g. *Echidna flea, Armored train*). 31% of cases

3. If a new child has an equivalent category, it is a Collection (*fallible, e.g. Category: New Zealand*). 8% of cases

4. If it has an infobox, looking at the relations in the infobox (e.g. *birthDate* applies to Individuals, *Kingdom* applies to Collections). 41% of cases

5. Unassigned are defaulted as individuals. 13%

Heuristics 2 and 3 were taken from work done at European Media Lab Research (Zirn et al, 2008)

Quality control is provided via Cyc's common-sense knowledge, as Cyc knows enough now to 'regurgitate' many assertions which are ontologically incorrect.

Examples of regurgitated assertions:

(#$isa #$CallumRoberts #$Research) ☹

Why it happened

Professor **Callum Roberts** is a marine conservation biologist, oceanographer, author and research scholar in the Environment Department of the University of York in England.

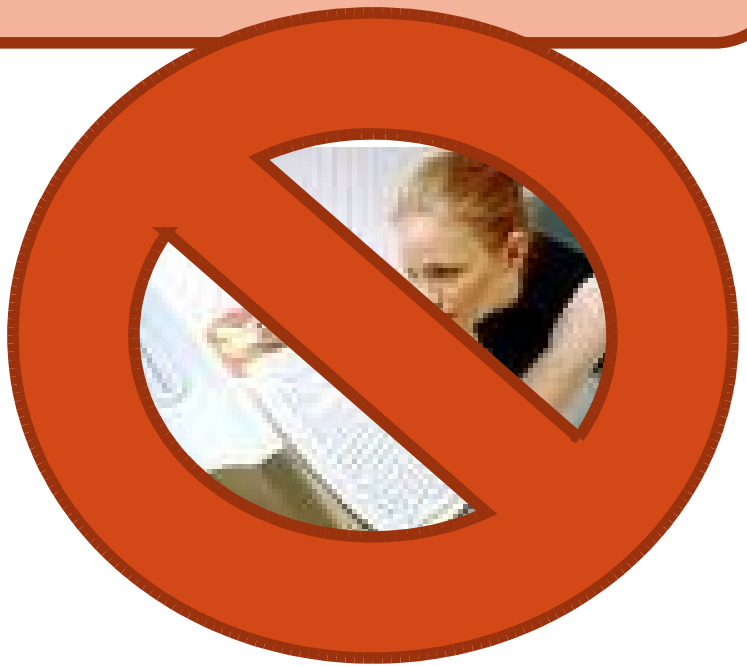
(#$isa #$Insight-EMailClient #$EMailMessage) ☹

Why it happened

Insight WebClient is an groupware E-Mail client from Bynari embedded on Arachne Web Browser for DOS.

```
BillGates is known not to be an instance of
      ParkingMeter in mt WikipediaToCycDataMt.
sbhl conflict: (isa BillGates ParkingMeter) TRUE
               WikipediaToCycDataMt
      because: (isa BillGates MaleHuman) True
               JustificationTruth
(genls MaleHuman MaleAnimal) TRUE
(genls MaleAnimal Animal) TRUE
(genls Animal AnimalBLO) TRUE
(genls AnimalBLO BiologicalLivingObject) TRUE
(disjointWith BiologicalLivingObject
      Artifact-Generic) TRUE
(genls Technology-Artifact Artifact-Generic) TRUE
(genls MechanicalDevice Technology-Artifact) TRUE
(genls ParkingMeter MechanicalDevice) TRUE
```

Bottom line: We have managed to add over 35K new concepts to Cyc entirely automatically! No ontological engineers involved!



Bootstrapping

- Our Stage C makes use of the ontology surrounding a given Cyc term to perform semantic disambiguation when finding mappings and new children.
- It follows that adding to the Cyc ontology might make further semantic disambiguation and yet further mappings and children possible.
- We tested this hypothesis by running our algorithm again, on a subset (10%) of the enlarged Cyc ontology, and were delighted to derive 1661 entirely new children.
- This extrapolates to an estimated 16K new children across the whole Cyc (46% of the size of the set derived by the first running of the algorithm).
- Achieving bootstrapping of a system's understanding is a long-held goal within AI research.

Evaluation:

- We used an online form to evaluate both the mappings and the new children created by the algorithm.
- 22 human volunteers participated in the evaluation, each answering at least 100 questions.
- We compared the new mappings against the 2008 mappings as a baseline.
- We compared our new children against parent- child pairs taken randomly from the DBpedia ontology.

CASE:	1	2	3	4	5	6
DBpedia children	0.58	0.81	0.99	0.98	0.99	0.99
New children	0.57	0.88	0.99	0.90	0.90	1.00
Old mappings	0.65	0.83	0.99	0.99	0.99	1.00
New mappings	0.68	0.91	1.00	1.00	1.00	1.00

CASES:

1 : 100% of evaluators thought assignment correct

2 : >50% thought assignment correct

3 : At least 1 thought assignment correct

4 : 100% thought assignment correct or close

5 : >50% thought assignment correct or close

6 : At least 1 thought assignment correct or close

The new mappings improved significantly on (Medelyan and Legg, 2008). Unfortunately the DBPedia ontology performed better, but it is manually checked, and also its assignments are much less specific than our algorithm

We make our results freely available at:

http://wdm.cs.waikato.ac.nz/cyc/portal/