Web Scale Reasoning and the LarKC Project (Review and Progress)

Michael Witbrock
Cycorp Europe
(with input from LarKC consortium members)
Goals of LarKC

LarKC = Large Knowledge Collider

- Build an integrated pluggable platform for large scale reasoning
  - Support for parallelization, distribution, remote execution, data storage
  - Use existing plug-ins, develop new
  - Easy integration of components
  - Enables low cost experimentation

“Significant progress is sometimes made not by making something possible that was impossible before, but by **substantially lowering the costs** of something that was only possible before at high cost”
Semantics
Commercially Relevant
Data Integration
Inference
Storage
Innovation was correct

Semantics

Commerically Relevant
Overall approach of LarKC

- Powerful platform with convenient service access
  - communication, synchronization, registration
  - LarKC = “SPARQL endpoint on steroids”
- Most work happens in the plugins
- LarKC gives you:
  - very scalable data layer
  - standardised interfaces for combining components
  - utilities & infrastructure to abstract from remote deployment
  - Inference about plugins

- Three types of LarKC users:
  - people building plugins
  - people configuring workflows
  - people using workflows
What do we mean by:

- **reusable** components
- **reconfigurable** workflows
- provide **infrastructure** needed by all users:
  - storage & retrieval
  - registration of plugins
  - communication (plugin2datalayer, plugin2plugins)
  - synchronisation (anytime behaviour)
  - remote execution (abstracts from local/remote storage)
  - remote data-access (abstracts from local/remote invocation)
  - (will) provide instrumentation & measuring
  - (will) provide caching & data-locality
- integration of very **heterogeneous** components
  - heterogeneous data: unstructured text, (semi)structured data
  - heterogeneous code: Java, scripts, remote services ("wrap & integrate")

LarKC = a **platform** for large scale reasoning
Goals of LarKC, and where we are

- Scalable: $> 10^9$ triples, lazy pipes
- Reconfigurable: plugins with standard API’s
- Open: Apache license
- Heterogenous: TRANSFORM, wrappers
- Experimentation: wrap & integrate
- Allow incompleteness: IDENTIFY, SELECT
- Enable distribution: plugin containers
- Anytime behaviour: streaming APIs
- Web-enabled: remote plugins & data
What Has Changed Since LarKC Started

• Scale is happening, and is important
  – academic work by Hitzler et al, Hendler et al, UT Dallas etc
  – commercial: OpenLink, Oracle, OntoText, BigData, …

• LOD accelerated growth and use (especially notable; Linked Govt Data)
The LarKC Platform Architecture 2.0

1) Start

2) Define Workflow

3) Pass workflow parameters

4) Workflow parameters

5) Instantiate, pass workflow parameters, notify user after successful setup

6) Query

LarKC entry point
- instantiates plug-in registry
- instantiates management interface (RDF endpoint)
- provides an executor per workflow (a single workflow may have multiple endpoints)

Management Interface

Executor

Endpoint (e.g., SPARQL)

User

Results

Transformer

Identifier

Selector

Decider

Reasoner

Decider

Backward chaining of plugin managers
What does a workflow look like?
What does a workflow look like?

Decider

Plug-in API

Plug-in Manager

Query Transformer

Identifer

Info Set Transformer

Selecter

Reasoner

Plug-in Registry

Plug-in Manager

Info Set Transformer

Selecter

Reasoner

RDF Store

Workflow Support System

Plug-in Manager

Query Transformer

Identifer

Info Set Transformer

Selecter

Reasoner
Decider Using Plug-in Registry to Create Workflow

Represent Properties
- Functional
- Non-functional (e.g. QoS)
- WSMO-Lite Syntax

Logical Representation
- Describes role
- Describes Inputs/Outputs
- Automatically extracted using API
- Decider can use for dynamic configuration
  - Rule-based
  - Fast
Example workflow

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<vbr:sparg xmlns:vbr="http://www.w3.org/2005/sparql-results#">
  <vbr:head>
    <vbr:variable vbr:name="COMPANY"/>
  </vbr:head>
  <vbr:results>
    <vbr:result>
      <vbr:binding vbr:name="COMPANY">
        <vbr:uri>http://www.cycfoundation.org/concepts/MicrosoftInc</vbr:uri>
      </vbr:binding>
    </vbr:result>
    <vbr:result>
      <vbr:binding vbr:name="COMPANY">
        <vbr:uri>http://www.cycfoundation.org/concepts/FordMotors</vbr:uri>
      </vbr:binding>
    </vbr:result>
  </vbr:results>
</vbr:sparg>
```
WP5: The LarKC Platform

open & flexible

modular & extendable

LarKC@SourceForge
LarKC Codebase
Plug-In API, Annotations and Plug-in Registry
Documentation
Realised LarKC Platform

Plugin Market
23 Plug-ins
LarKC Data Layer

Application

- Plug-in API
- Decider
- Pipeline Support System
- Plug-in Registry
- Plug-in Manager
  - Plug-in API
  - Query Transformer
- Plug-in Manager
  - Plug-in API
  - Identifier
- Plug-in Manager
  - Plug-in API
  - Info. Set Transformer
- Plug-in Manager
  - Plug-in API
  - Selecter
- Plug-in Manager
  - Plug-in API
  - Reasoner

Data Layer API

Data Layer

- RDF Store
- RDF Store
- RDF Store
- RDF Doc
- RDF Doc
- RDF Doc
Main goal:

- The Data Layer supports LarKC plug-ins:
  - storage, retrieval and light-weight inference on top of large volumes of data
  - automates the exchange of RDF data by reference and by value
  - offers other utility tools to manage data (e.g. merger)
Used Concepts in the Data Model

Labelled groups of statements

NG1
NG2
NG3
NG4
NG5
## Supported Sets of Statements

<table>
<thead>
<tr>
<th>RDF data types</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set of statement</td>
<td>RDF statements</td>
<td>s1, p1, o1, ng1, s2, p2, o2, s3, p3, o3, ng3, {group1}</td>
</tr>
<tr>
<td>RDF graph</td>
<td>Named graph</td>
<td>s1, p1, o1, ng1, {group1}, s2, p2, o2, ng1, {group2}, s3, p3, o3, ng1</td>
</tr>
<tr>
<td>Dataset</td>
<td>SPARQL dataset</td>
<td>s1, p1, o1, ng1, s2, p2, o2, ng2, s3, p3, o3, ng3</td>
</tr>
<tr>
<td>Labelled group of statements</td>
<td>RDF group of statements</td>
<td>s1, p1, o1, ng1, {group1}, s2, p2, o2, ng2, {group1}, s3, p3, o3, ng3</td>
</tr>
</tbody>
</table>
### Different Implementations

<table>
<thead>
<tr>
<th>Name</th>
<th>Purpose</th>
<th>Cost to pass $10^6$ RDF statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set of statements</td>
<td>Services to return RDF data (e.g., SPARQL construct and describe queries)</td>
<td>350MB memory</td>
</tr>
<tr>
<td>Named graph in memory</td>
<td>Pass named graphs between plug-ins</td>
<td>350MB memory</td>
</tr>
<tr>
<td>HTTP published RDF graph</td>
<td>Load chunks of RDF from URL</td>
<td>few KBs + time to load the data</td>
</tr>
<tr>
<td>Named graph retrieved from a SPARQL endpoint</td>
<td>Pass named graphs using a SPARQL endpoint</td>
<td>few KBs + time to transfer the data + to filter the dataset with SPARQL</td>
</tr>
<tr>
<td>SPARQL dataset</td>
<td>Constrain RDF data exposed by a SPARQL endpoint with a list of graph names</td>
<td>few KBs + time to transfer the data + to filter the dataset with SPARQL</td>
</tr>
<tr>
<td>Labelled group of statements</td>
<td>Pass arbitrary sets of RDF statements between plug-ins (e.g., graph priming)</td>
<td>few KBs + time to associate the statements + time to filter the labelled group</td>
</tr>
</tbody>
</table>
WP11: Instrumentation
Released System v1.1: larkc.sourceforge.net

- Open Apache 2.0 license
- Previous early adopters workshops @ ESWC ’09,10 and ISWC ‘09
  - participants modified plug-ins, modified workflows

Standard Open Environment:
Java, subversion, packaged release, command line build, or eclipse
Research Topics
Part of broader effort led by Ontotext, WICI (Beijing), Max Planck (Berlin) to apply techniques derived from human psychology to web-scale reasoning.

- The performance of the spreading activation tasks varies considerably depending on the parameters of the process.

- As a reference point use the following result: it takes **7 seconds to activate about 7 thousand nodes** after spreading of activation from resource [http://dbpedia.org/resource/Berlin](http://dbpedia.org/resource/Berlin) with decay factor 0.25.

- Queries on the “primed” or “selected” part of a dataset run up to 20 times faster and return only focussed results.

<table>
<thead>
<tr>
<th>Query performance after selection through priming</th>
<th>Query 1:</th>
<th>Query 2</th>
</tr>
</thead>
<tbody>
<tr>
<td># of active nodes after graph priming</td>
<td>368,381</td>
<td>368,413</td>
</tr>
<tr>
<td># of results (time for eval) before priming</td>
<td>2,174 (87 ms)</td>
<td>163,438 (518 ms)</td>
</tr>
<tr>
<td># of results (time for eval) after priming</td>
<td>23 (5 ms)</td>
<td>530 (47 ms)</td>
</tr>
<tr>
<td>Reduction ratio for the number of results</td>
<td>94.5</td>
<td>308.4</td>
</tr>
</tbody>
</table>
Achieved: experimental setup for applying IR, Cognitive Memory and Spreading Activation methods for the problem of selection/subsetting/RDF search on large scale

Demonstrated:
- Scalable, parallelised selection strategies:
  - Cognitive Memory:
    - Interests-based selection use cases for medical researchers and computer scientists
  - Sub-setting based on statistical semantics applied to text and then linked to RDF Graphs
  - Statistical semantics applied directly to large RDF graphs (RDF molecules)
- Cutting-edge work on parallelisation
  - Cognitive Memory Spreading activation demonstrated on CUDA “GPU” cards
Data Streams are Ubiquitous

- “Is a traffic jam going to happen in this highway? And is then convenient to reallocate travelers based upon the forecast?”
- “By looking at the click stream coming from a given IP, can we notice the shifts of interest of the person behind the computer?”
- “Which contents of the news Web portal are attracting more attention? Which navigation pattern would lead readers to other news related to those contents?”
- “Are trends in medical records indicative of any new disease spreading in given parts of the world?”
- “Where are all my friends meeting?”
- “In the financial context, can we detect any intraday correlation clusters among stock exchange?”
C-SPARQL extends SPARQL to support continuous queries,
queries registered over RDF data streams and then continuously executed.
Queries consider *windows*, i.e. the most recent triples of such streams, observed while data is continuously flowing.
WP3: Abstraction and Learning

Machine learning for the Semantic Web

About: Angela Merkel
An entity in dataspace: dbpedia.org

Predicates that are predicated are highlighted. Click on the more predictions links next to the predicate names to see more predictions.

- dboedia2:order
  - more predictions
  - Chancellor of Germany
- dboedia2:occupation
  - more predictions
  - Lawyer
- dbpedia:ontology/Person
  - more predictions
  - Politician
- dbpedia:ontology/occupation
  - more predictions
  - Politician

Traffic-Aware Routing Showcase

Social Media Analytics using C-SPARQL and SUNS

Legend
- data stream
- RDF stream
- RDF graph
- C-SPARQL query
- SPARQL with Probability
- Inductive Reasoner
- Long-Term Matrix
- Abstracter
- Deductive Reasoner
- Social Media Analytics
WP 4: Reasoning and Deciding - WebPIE

- Uses Cluster/MapReduce to distribute computation
- RDFS and the OWL-horst rule-sets
- Full forward inference over large datasets

Reasoning over Life Data
Use transitivity to find all types of cancer (long transitive chains)

- First in SemWeb community (to break the 100B triple barrier)
- Now on Amazon EC2 for push-button use by 3rd parties.
MaRVIN Architecture

• MaRVIN is:
  – a platform for distributed RDF(S) reasoning
  – a platform for processing lots of RDF data

• MaRVIN scales by:
  – distributing computation over many nodes
  – approximate (sound but incomplete) reasoning
  – anytime convergence (more complete over time)
WP5: The LarKC Platform

LarKC@SourceForge

Plug-In API, Annotations and Plug-in Registry

Documentation

Realised LarKC Platform

open & flexible

modular & extendable

LarKC Codebase

Plugin Market

23 Plug-ins
WP5: Parallelisation Patterns

Transformer Plug-in (WP6) Speed-up: ≈ 4

MPI parallelisation of Random Indexing code (WP2,7) Speed-up: ≈ 20-40

Plug-in parallelisation

- Multi-threading
- MPI
- MapReduce
WP5: Platform Distributed Execution Support

Decider

queue control system

Plug-in Manager

Plug-in API

Identifier

Plug-in Manager

Plug-in API

Remote Launch Mechanism

Plug-in Manager

Plug-in API

Selector

Plug-in Manager

Plug-in API

Identifier

http://

ssh://

powered by JEE technology

powered by GAT

powered by JEE technology

powered by GAT
WP6 - Urban Computing Use Case

Practical: Involvement of Korean Institute of Building Technology joint with SaltLux. Significant datasets (1B triples) in realistic case-study aimed at Korean govt.

Road Sign Management Demonstrator
Demonstrates Reasoning (WP4)

Traffic-Aware Routing Showcase
Demonstrates Abstraction & Learning (WP3)

Journal Publication
Linking and Mapping Semantics for Early Clinical Development

Number of Statements: 4,193,400,044
- Explicit statements: 2,741,704,569
- Inferred statements: 1,451,695,475

Cross dataset mappings: 2,156,645
Semantic Annotations: 263,323,164
Number of RDF resources: 582,691,283
Size on disk: ~320GB
Unique users Jan-Aug 2010: 10,226

Asthma and chronic obstructive pulmonary disease (COPD) are chronic airway diseases characterized by airflow obstruction. The beta(2)-adrenoceptor mediates bronchodilation in response to exogenous and endogenous beta-adrenoceptor agonists. Single nucleotide polymorphisms in the beta(2)-adrenoceptor gene (ADRB2) cause amino acid changes (e.g., Arg16Gly, Gln27Glu) that potentially alter receptor function.
Example Query

- Integrates Multiple Knowledge Sources and Inputs

Give me all proteins which interact in cellular structure and are annotated with repressor and have at least one participant that is encoded by a gene annotated with specific term and is located in chromosome X? Filter the results for Mammalia organisms!
WP 7a Linked Data used through LarKC

• Interest Based Selection (WP2)
  – Refine the results based on who you are
  – Provide a better selection based on who you are

• Random Indexing Selection (WP2)
  – Implement generic RDF search and selection
  – Assist the user in exploring datasets with uncertain schemata

• Causality Chain Mining for Clinical Discovery (WP3)
  – Explore possible causative chains between heterogeneous data sources
  – Rank the results using feedback mechanism
Initial study (last year) included 650k SNPs genotyped in 1229 cases and 2450 controls – study too small and no apparent findings.

- Re-ranked SNPs based on LarKC priors - top ranked SNP now rs7579899 (originally ranked 28th).
- Subsequently, study data pooled with multiple GWAS from Europe and US (one year worth of work + $1M)
  rs7579899 among top 5 ranked SNPs

<table>
<thead>
<tr>
<th>Study</th>
<th>cases</th>
<th>controls</th>
<th>OR</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial study</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CE</td>
<td>1096</td>
<td>2062</td>
<td>1.19</td>
<td>1.07-1.33</td>
</tr>
<tr>
<td>HUNT2/Tromso</td>
<td>133</td>
<td>388</td>
<td>1.07</td>
<td>0.80-1.42</td>
</tr>
<tr>
<td>Further studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASHRAM</td>
<td>84</td>
<td>399</td>
<td>1.32</td>
<td>0.87-1.99</td>
</tr>
<tr>
<td>EPIC</td>
<td>276</td>
<td>419</td>
<td>1.12</td>
<td>0.88-1.42</td>
</tr>
<tr>
<td>CeRePP</td>
<td>79</td>
<td>146</td>
<td>1.38</td>
<td>0.92-2.05</td>
</tr>
<tr>
<td>Moscow</td>
<td>263</td>
<td>306</td>
<td>1.28</td>
<td>1.00-1.64</td>
</tr>
<tr>
<td>SEARCH</td>
<td>530</td>
<td>1361</td>
<td>1.15</td>
<td>0.99-1.34</td>
</tr>
<tr>
<td>ATBC</td>
<td>163</td>
<td>1288</td>
<td>1.37</td>
<td>1.09-1.73</td>
</tr>
<tr>
<td>CPSII</td>
<td>207</td>
<td>730</td>
<td>1.32</td>
<td>1.05-1.65</td>
</tr>
<tr>
<td>PLCO</td>
<td>284</td>
<td>845</td>
<td>1.05</td>
<td>0.87-1.28</td>
</tr>
<tr>
<td>USKC</td>
<td>662</td>
<td>561</td>
<td>1.11</td>
<td>0.94-1.31</td>
</tr>
</tbody>
</table>

- rs7579899 is located in the *EPAS1* gene which is involved in several mechanisms relevant to kidney cancer.
- A report outlining these results has been tentatively accepted for publication in Nature Genetics (Purdue & Johansson et al. 2010)
Mímir: Simultaneous Semantic and Text search

{Gene} is implicated in {Disease} of {partOf Lung}
• **Human Computer Collaboration**
  ◦ Vast power: fast communications, near infinite memory, computational learning (recommendation, data mining)

• **Fundamentally Harder Problems**
  ◦ This power is creating systems that are more complex than can be managed by unaided humans

• **Fundamental Change**
  ◦ Much content not stored in human understandable form (text, audio, diagrams ...) but dynamically combined and synthesized

• **Needs infrastructure**
  ◦ Knowledge in machine understandable form
  ◦ Background knowledge to provide that understanding
Use Case: Semantic Patient Records

Reasoning-based, question answering

User-assisted query understanding
In use at the Cleveland Clinic (one of largest US hospitals)
Content adapted to query: Heart Valve repairs by Dr. Gillinov

Content adaptation: heart valve repair
A. Marc Gillinov, M.D.

Gillinov, M.D. has performed 316 coronary artery operations at the Cleveland Clinic Foundation since 1998. He specializes in minimally invasive mitral valve, aortic valve, and tricuspid valve surgery; mitral valve repair; surgical treatment and minimally invasive surgery for atrial fibrillation; off-pump coronary artery bypass surgery; and high-risk mitral valve surgery.

**Medical Degree:** Johns Hopkins University School of Medicine, Baltimore, Maryland

**Special Training:** Johns Hopkins University School of Medicine, Baltimore, Maryland

A native Clevelandian, Dr. Gillinov spent the summers of 1976 to 1980 dividing his time between working at Cleveland Clinic and playing tennis. He still enjoys both activities.

Content adapted to new query: **Artery** repairs by Dr. Gillinov
Practical: In use with live records

Standards:
- Patient Records are in triple Store
- Records accessed via standard SPARQL Queries

Cyc adds:
- Background Medical Knowledge
- Natural Language Interface
- More Powerful Inference over Standard Data

API Support:
- Web Service APIs
- Java APIs
OpenCyc Collection: shipping container

A specialization of StorageConstruct. Each instance of LargePortableCargoContainer is a large (usually about 40' x 8' x 8') metal container used for shipping large quantities of goods across land (e.g., on a large truck) or overseas (e.g., on a cargo ship).

A Type of: storage device

Instance of: spatially disjoint object type, type of object

Subtypes:

Instances:

Same as:

http://umbel.org/umbel/sc/LargePortableCargoContainer
OpenCyc Collection: loading a vehicle

Unique ID: [ Mx4rSbwOIPpYEdqAAAACs71DGQ ]
English ID: [ LoadingAVehicle ]
English Aliases: [ "had loaded a vehicle", "has loaded a vehicle", "have loaded a vehicle", "load a vehicle", "loaded a vehicle", "loads a vehicle", "will have loaded a vehicle", "will load a vehicle" ]

The collection of instances of Loading in which the toLocation is a vehicle.

A Type of: loading, putting something into something else
Instance of: type of temporally stuff-like thing
Subtypes:
Instances:
Same as:
OpenCyc Individual: South Korea

OpenCyc Individual: Bay of Korea

OpenCyc Individual: Korean script

OpenCyc Collection: Korean food

OpenCyc Collection: ginseng

The collection of all ginseng. A type of UmbelliferOrder, DrugSubstance, and Plant_NonWoody. Ginseng is a MedicinalPlantType, a SpatiallyDisjointObjectType, and an OrganismClassificationType.

A Type of: drug, herbaceous plant, umbilifer order

Subtypes:

Instances:


Same as:

http://umbel.org/umbel/sc/Ginseng
http://dbpedia.org/resource/Ginseng
03/02/2010
10:34:19 (relationInstanceAll nationOfState SouthKorea Province-SouthKorea) by DaveS

03/13/2009
18:50:21 (broaderTerm CityOfSeoulSouthKorea SouthKorea) by Blake

03/12/2009
21:23:07 (rdf-type SouthKorea IndependentCountry) by Blake
21:22:48 (broaderTerm SouthKorea IndependentCountry) by Blake

09/03/1997
(capitalCity SouthKorea CityOfSeoulSouthKorea) by Foxvog

10/03/1996
(nameString SouthKorea "Korea") by KathyPanton

04/03/1996
(codeMapping ISOCountryDigraphMapping "KR" SouthKorea) by Colvin
Republic of the Congo is a member of Organization of African Unity, the United Nations, and African Union.
Seoul, the Republic of Korea

Seoul, Korea, Republic of is a city in South Korea and a capital.

Seoul is the capital of South Korea. Seoul, Korea, Republic of is in South Korea.

Seoul, Korea, Republic of's latitude is 37.56639862060547 degrees. Seoul, Korea, Republic of's longitude is 127 degrees.

Miscellaneous

Some subway is located in Seoul, Korea, Republic of. In 1994, Seoul, Korea, Republic of had a population of 9,204,344 people. Seoul, Korea, Republic of's population was 9,204,344 in 1994.

Manually Written Notes

The #capitalCity of South Korea.

the CIA World Factbook: 1995 edition

Technical Details

This page was automatically created by Cyc using automated reasoning and natural language processing.

Cyc first learned about Seoul, the Republic of Korea on November 29, 1987. The Semantic Web URI for Seoul, the Republic of Korea is: http://sw.opencyc.org/concept/Mx4rvVjXhJwpEbGdrcN5Y29ycA

and an East Asian country.

- South Korea is west of Japan.
- China is west of South Korea.
People

- \( \text{isa} \) occupation BioDataPredicate
- \( \text{isa} \) dateOfDeathEvent BioDataPredicate
- \( \text{isa} \) placeOfDeath BioDataPredicate
- \( \text{isa} \) citizens BioDataPredicate
- \( \text{isa} \) ethnicity BioDataPredicate
- \( \text{isa} \) birthPlace BioDataPredicate
- \( \text{isa} \) dateOfDeath BioDataPredicate
- \( \text{isa} \) birthDate BioDataPredicate

Agent Location

\( \text{Mt} : \) BaseKB
\( \text{Direction} : \) Backward

\( \text{M(implies)} \)

\( \text{and} \)

- \( \text{genlPreds} \) ?SPEC affiliatedWith
- \( \text{ist-Asserted} \) ?MT
  - \( ?\text{SPEC} \) ?ARG1 ?ARG2
- \( \text{agentAffiliationDataSentence} \) ?ARG2
  - \( \text{ist} \) ?MT
  - \( ?\text{SPEC} \) ?ARG1 ?ARG2

Agent Affiliation

\( \text{M(implies)} \)

\( \text{and} \)

- \( \text{genlPreds} \) ?SPEC affiliatedWith
- \( \text{ist-Asserted} \) ?MT
  - \( ?\text{SPEC} \) ?ARG1 ?ARG2
- \( \text{agentAffiliationDataSentence} \) ?ARG1
  - \( \text{ist} \) ?MT
  - \( ?\text{SPEC} \) ?ARG1 ?ARG2

What goes on a inCyc Page?
\textbf{InCyc HTML Production Rules}

\begin{verbatim}
M\text{implies}
  (and
    (?PRED BasicDataPredicate)
    (?miniFactSheetFlagHtml ?TOPIC ?FLAG-IMAGE-HTML)
    (?PRED ?TOPIC ?DATA)
    (evaluate ?FLAG-DATUM
      (HTMLTableDataFn
        (TheList
          (TheList Rowspan-HTMLTagAttribute 3)) ?FLAG-IMAGE-HTML))
    (evaluate ?ROW
      (HTMLTableRowFn TheEmptyList ?FLAG-DATUM ?ROW1))
    (evaluate ?ROW1
      (HTMLTableDataFn TheEmptyList
        (BasicDatumParaphraseFn
          (NthInListFn ?DATA 1) ?MT)))
    (evaluate ?ROW2
      (HTMLTableRowFn TheEmptyList
        (HTMLTableDataFn TheEmptyList
          (BasicDatumParaphraseFn
            (NthInListFn ?DATA 2) ?MT))))
    (evaluate ?ROW3
      (HTMLTableRowFn TheEmptyList
        (HTMLTableDataFn TheEmptyList
          (BasicDatumParaphraseFn
            (NthInListFn ?DATA 3) ?MT))))
    (evaluate ?TABLE
      (HTMLTableFn TheEmptyList
        ?ROW ?ROW2 ?ROW3))
    (basicDataHTML ?TOPIC ?MT ?TABLE))
\end{verbatim}
Cyc contains:
- 15,000 Predicates
- 300,000 Concepts
- 3,200,000 Assertions

Represented in:
- First Order Logic
- Higher Order Logic
- Modal Logic
- Context Logic
- Micro-theories

General Knowledge about Various Domains

Specific data, facts, and observations
Very specific information
(some indirect, via SKSI)

Upper Ontology

Core Theories

Domain-Specific Theories

EVENT \supset \text{TEMPORAL-THING} \supset \text{PARTIALLY-TANGIBLE-THING}

\(( \forall a, b ) \ a \in \text{EVENT} \land b \in \text{EVENT} \Rightarrow \text{causes}( a, b ) \Rightarrow \text{precedes}( a, b )\)

\(( \forall m, a ) m \in \text{MAMMAL} \land a \in \text{ANTHRAX} \Rightarrow \text{causes( exposed-to( m, a ), infected-by( m, a ) )}\)

Very specific information (some indirect, via SKSI)

First Order Predicate Calculus: unambiguous; enable mechanical reasoning

\[
\exists y. \forall x. \text{Amer}(x) \Rightarrow \text{president}(x,y) \\
\forall x. \exists y. \text{Amer}(x) \Rightarrow \text{mother}(x,y)
\]

Higher Order Logic: contexts, predicates as variables, nested modals, reflection,...

Formal Representation for Reasoning
First Order

• (isa ASBFinancialCorp PubliclyHeldCorporation)
• (corporateOfficers ASBFinancialCorp GeraldRJenkins)

With Context

• In Mt: FinancialTransactionMt
  (relationAllExists performedBy RepurchaseProgram PubliclyHeldCorporation)

Rule

• In Mt: FinancialTransactionMt
  (forAll ?X  (implies
     (isa ?X RepurchaseProgram)
     (thereExists ?Y (and (isa ?Y PublicallyHeldCorporation) (performedBy ?X ?Y))))))

Second Order

• (implies
  (equals ?SET (TheSet ?THING)))

Modal

• (beliefs Israel (relationInstanceExists possesses Syria ClusterBomb))

Meta

• (opaqueArgument beliefs 2)
Is it attached to the inside of the outer object?
- Yes -- Try #$connectedToInside

Can it be removed by pulling, if enough force is used, without damaging either object?
- No -- Try #$in-Snugly or #$screwedIn

Does the inner object stick into the outer object?
- Yes -- Try #$sticksInto
Does part of the inner object stick out of the container?

- None of it. \#$in-ContCompletely
- Yes \#$in-ContPartially
- No \#$in-ContClosed

If the container were turned around could the contained object fall out?

- Yes \#$in-ContOpen

For Inference: Senses of ‘In’
Some Transportation Event Types

- "$TransportationEvent"
- "$ControllingATransportationDevice"
- "$TransportWithMotorizedLandVehicle"
  - ("$SteeringFn $RoadVehicle")
- "$TransporterCrashEvent"
- "$VehicleAccident"
- "$CarAccident"
- "$Colliding"
- "$IncurringDamage"
- "$TippingOver"
- "$Navigating"
- "$EnteringAVehicle ..."
Web 3.0 Systems start from Web 2.0-style learning.
Acquire ground facts, test rule inferences.

Knowledge Acquisition with CURE
Palm Inc is a publicly held corporation with its headquarters in Sunnyvale, California. It was founded by Ed Colligan and operates in the high-tech manufacturing enterprise sector.
Renaissance Artists

Bronze Age Farmers

Kind of TimeInterval

Noun Form: not plural

Kind of Agent-Generic

Noun form

(SubcollectionOfWithRelationToFn Artist activeDuringPeriod TheRenaissance)

(SubcollectionOfWithRelationToFn Farmer activeDuringPeriod TheBronzeAge)
Attributive Adjective & Noun

Yellow Submarine
(SubcollectionWithRelationToFn Submarine
mainColorOfObject Yellow)

Sick Child
(SubcollectionWithRelationToFn HumanChild
stateOfHealth Sick)

Supporting Lexical Assertions:

(adjSemTrans Yellow-TheWord 0 RegularAdjFrame
(mainColorOfObject :NOUN YellowColor))

(adjSemTrans Sick-TheWord 0 RegularAdjFrame
(stateOfHealth :NOUN Sick))
Brussels neighborhood is a Brussels neighborhood.
Brussels neighborhood

Advanced Info

Bruxelles Centre Ville is a Brussels neighborhood.

Not Understood

Unable to understand "Bruxelles Centre Ville".
Would you like to define it?

Yes  No
Brussels neighborhood

About  Basic Info  Similar  Advanced Info

Bruxelles Centre Ville  is a Brussels neighborhood.

Not Understood

Which of the following describe 'Bruxelles Centre Ville'?

- Brussels neighborhood

Create  Cancel
Brussels neighborhood

- About
- Basic Info
- Similar
- Advanced Info

Bruxelles Centre Ville is a Brussels neighborhood.
Bruxelles Centre Ville

Bruxelles Centre Ville properly geographically subsumes La Grande Place.

Bruxelles Centre Ville.

Someone should visit the cafe when visiting Bruxelles Centre Ville.

Someone should visit the restaurant space when visiting Bruxelles Centre Ville.

Localized event happened in Bruxelles Centre Ville.
Main courses
(with choice)

Dish of the day

or Tartar steak, frites

or Stewed beef with gueuze sauce and frites

or Grilled salmon béarnaise

or Mussels in casserole, frites (in season, 600g)
Bruxelles Centre Ville properly geographically subsumes La Grande Place.

Brussels, Belgium properly geographically subsumes Bruxelles Centre Ville.

Someone should visit the cafe when visiting Bruxelles Centre Ville.

Which of the following describe 'Aux Armes de Bruxelles'?
- restaurant space

Create | Cancel
Aux Armes de Bruxelles has **type of dish** on the menu.

Aux Armes de Bruxelles serves **cuisine**.

**chef** is the chef at Aux Armes de Bruxelles.

**sous-chef** is sous-chef at Aux Armes de Bruxelles.

**waitperson** is waitperson at Aux Armes de Bruxelles.

Each shift at Aux Armes de Bruxelles have **non-negative integer** people working.

Work happens during **calendar day** from **time of day type** to **time of day type** in Aux Armes de Bruxelles.

**localized event** happens on **Choose a value** from **time of day type** to **time of day type** at Aux Armes de Bruxelles.

The last part of the wikipedia URL for Aux Armes de Bruxelles is.

Aux Armes de Bruxelles is located in **spatial thing non situational**.
Aux Armes de Bruxelles has type of dish on the menu.

Aux Armes de Bruxelles serves Belgian cuisine.

Aux Armes de Bruxelles serves chef is the chef at Aux Armes de Bruxelles.

Aux Armes de Bruxelles serves sous-chef is sous-chef at Aux Armes de Bruxelles.

Aux Armes de Bruxelles serves waitperson is waitperson at Aux Armes de Bruxelles.

Each shift at Aux Armes de Bruxelles have non-negative integer people working.

Work happens during calendar day from time of day type to time of day type in Aux Armes de Bruxelles.

Localized event happens on time of day type to time of day type at Aux Armes de Bruxelles.

The last part of the wikipedia URL for Aux Armes de Bruxelles is.

Aux Armes de Bruxelles is located in spatial thing non situational.
Michael Witbrock

Michael Witbrock's occupation is artificial intelligence researcher.
Michael Witbrock's occupation is computer scientist.
Michael Witbrock resides in Austin.
Michael Witbrock was born in Christchurch, New Zealand.
Michael Witbrock was born in New Zealand.
Michael Witbrock was born on September 17, 1962.
Demo Movie: Knowledge Use

Access ground facts, apply rule inferences.

Knowledge Use with CAE